Forsythe, R. O., Ozdemir, B. A., Chemla, E. S., Jones, K. G., & Hinchliffe, R. J. (2016). Interobserver Reliability of Three Validated Scoring Systems in the Assessment of Diabetic Foot Ulcers. *International Journal of Lower Extremity Wounds*, *15*(3), 213-219. https://doi.org/10.1177/1534734616654567

ABSTRACT

Scoring systems for diabetic foot ulcers may be used for clinical, research or audit, to help assess disease severity, plan management and even predict outcomes. Whilst many have been validated in study populations, little is known about their inter-observer reliability. This prospective study aimed to evaluate inter-observer reliability of three scoring systems for diabetic foot ulceration.

After sharp debridement, diabetic foot ulcers were classified by a multi-disciplinary pool of trained observers, using the PEDIS (Perfusion, Extent, Depth, Infection, Sensation), SINBAD (Site, Ischaemia, Neuropathy, Bacterial infection, Depth) and University of Texas wound classification systems. Inter-observer reliability was assessed using intra-class correlations (0 = no agreement; 1 = complete agreement).

Some 37 patients (78.4% male) were assessed by a pool of 12 observers. Single observer reliability was slight to moderate for all scoring systems (UT 0.53; SINBAD 0.44, PEDIS 0.23-0.42) but multiple observer reliability was almost perfect (UT 0.94; SINBAD 0.91; PEDIS 0.80-0.90). The worst agreement for single observers was when scoring infection (SINBAD 0.28; PEDIS 0.28), ischaemia (SINBAD 0.26; PEDIS 0.23) or both (UT 0.25), however this improved to almost perfect agreement for multiple observers (infection: 0.83; ischaemia: 0.80-0.82; both: 0.81).

These classification systems may be reliably used by multiple observers, for example when conducting research and audit. However, they demonstrate only slight to moderate reliability when used by a single observer on an

individual subject and may therefore be less helpful in the clinical setting, when documenting ulcer characteristics or communicating between colleagues.

Diabetic foot ulcers (DFU) are a serious complication of diabetes, leading to significant morbidity. It is estimated that up to 15% of patients with diabetes may develop a foot ulcer during their lifetime [1] and the majority of patients requiring major lower limb amputation have had a preceding foot ulcer.

The heterogeneity of DFU disease progression and outcomes makes it difficult to apply population-based outcomes data to an individual patient. In addition, the individual factors such as peripheral artery disease (PAD), neuropathy and ulcer size, may influence the healing of a DFU by varying degrees and the interplay between these factors makes risk prediction challenging. In order to address these issues, a number of scoring and classification systems have been developed to aid clinicians when assessing DFU, [2] which vary in complexity. According to the IWGDF, the aim of a classification system for diabetic foot ulcers in clinical practice should be to facilitate communication between health professionals, influence daily management and provide information about the healing potential of an ulcer.[3] Details on the aetiology of the ulcer, as well as patient characteristics, are required in order to use a scoring system in audit or research, if the ultimate aim is to identify appropriate treatment strategies, evaluate disease prevalence or perform more complicated analyses such as exploring differences in outcomes between centres. In contrast, clinical scoring systems should be relatively simple, easy to use and may only require basic details on ulcer characteristics, in order to facilitate accurate documentation during clinical assessment or to allow tracking of lesions throughout an episode of care, which may allow improved communication between colleagues.

Whilst many classification systems have been appropriately validated internally and, in some cases, externally, inter-observer reliability (ie repeated measurements of a stable condition produces similar results when scored by different observers) has not been widely reported for most of the current scoring systems. A well-constructed and validated system will not be useful if it has poor inter-observer reliability. Evaluation of inter-observer reliability is also required in order to ensure that assimilation of data within and across multiple sites can be meaningfully interpreted. This may improve the power and quality of research studies but may also prove helpful when considering the use of validated scoring systems in clinical practice.

In this study, we aimed to determine the inter-observer reliability in the use of three well-known validated scoring systems for DFU, PEDIS[3] (Perfusion, Extent, Depth, Infection, Sensation), SINBAD[4] (Site, Ischaemia, Neuropathy, Bacterial infection, Depth) and University of Texas (UT)[5] wound classification systems.

Materials and methods

Study design

This was a prospective, single-centre observational study of patients already engaged in a multi-disciplinary diabetic foot clinic at a large teaching hospital. Approval was granted from a local research ethics committee prior to recruitment (NRES Committee London – Stanmore; ref 13/LO/1431) and the research was performed according to the Declaration of Helsinki (2008).

The study was undertaken during the participants' routine clinical visits. The usual standard of care was maintained throughout the study, according to national guidelines[6] and comprised input from a full range of health professionals including vascular surgeons, podiatrists, diabetologists, microbiologists, radiologists and orthopaedic surgeons. There was no alteration to the standard clinical care provided to the participants for the duration of the study.

Participants

Potential participants were approached by a member of the multi-disciplinary team during the weekly multi-disciplinary diabetic foot clinic and given verbal and written information about the study. After screening for and confirmation of eligibility, willing participants provided informed written informed consent prior to enrolment. Presence of a diabetic foot ulcer was defined as per the International Consensus on the Diabetic Foot as 'a full-thickness wound below the ankle in a diabetic patient, irrespective of duration, tissue necrosis and gangrene'. Inclusion criteria were: 1) presence of a diabetic foot ulcer; 2) Age over 18; 3) Known to, and being treated by, the diabetic foot service in participating centre; 4) has read the Patient Information Leaflet and given informed consent. Exclusion criteria were: 1) Unable to give informed consent; 2) Clinically too unwell to participate.

Clinician assessors

Participants were assessed by a pool of 12 multi-disciplinary health professionals usually involved in the care of patients with diabetes, including vascular surgeons, diabetologists and podiatrists. All were members of the

local diabetic foot team and had experience in treating and managing patients with DFU, including the palpation of foot pulses as part of the clinical assessment of these patients. Assessors received an introductory lecture prior to the study commencing and a summary of the use of the scoring systems was given on each day of assessment, as well as a demonstration of how to use the instruments. A vascular surgeon who was familiar with the use of the instruments supervised their use, however assessors were not individually tested on their ability to perform the procedures. The specialist diabetic foot clinic at this hospital contributes to the National Diabetes Foot Care Audit, which uses scoring systems (such as SINBAD) to report ulcer characteristics. The assessors were therefore familiar with their use.

Assessments and follow-up

Patients were assessed at a single clinic visit, following the completion of routine medical care including sharp tissue debridement by a podiatrist. Tissue debridement was considered the removal of non-viable and necrotic tissue and callus using a sharp instrument in the clinic setting, in order to promote wound healing. This did not include surgical or complex debridement. Demographic data was collected on each participant and the study assessment was performed by each observer separately and without collaboration. When the patients were evaluated by multiple observers, this was carried out during the same session. Each clinician assessor completed the clinical assessment according to a pre-prepared checklist and scores were calculated at the end of the assessment. Upon completion of the study assessment, participants continued their usual pathway of care; there was no additional follow-up required for the study. A pool of clinician assessors was

used, as it was not possible for the same assessors to assess every patient in the study, due to varying clinical commitments over the study period.

Data measurements

All clinicians used the same equipment throughout the duration of the study. Assessors were encouraged to ask patients questions to elicit symptoms of peripheral artery disease or infection but were not permitted to review results from objective tests such as duplex ultrasound. In addition, assessors were provided with a ruler, a 10g monofilament, 128 Hz tuning fork and probe to use during their examination. Assessment of perfusion was made by palpating the foot pulses. In order to satisfy criteria for the lowest grading of ischaemia on the PEDIS classification (Grade 1), it is permitted to use presence of both foot pulses (in addition to the absence of symptoms of PAD). However, if both foot pulses are not palpable, the scoring system requires the use of objective testing, using ankle brachial index (ABI), toe brachial index (TBI) or transcutaneous oxygen pressure (TcPO2). In this study, objective testing of perfusion was not carried out if foot pulses were found to be absent.

Data analysis

Study size and statistical analysis

Analysis was performed using an intra-class correlation (ICC) test, [7] which measures agreement and the overall data variance due to between-subjects variability, when the subjects are measured by a different sample of observers for each subject drawn from an infinite pool of observers. [8, 9] ICC (1,1) measures reliability of a single observer reporting on individual subjects,

whereas ICC (1,k) reports the reliability of multiple observers' average ratings for a group of subjects.

The UT and SINBAD scoring systems comprise a number of assessment domains, culminating in an overall aggregate score or category, whereas the PEDIS system assesses 5 domains and reports them separately, with no overall score (Table 1). Therefore, for the UT and SINBAD scoring systems, ICC (1,1) and ICC (1,k) were reported for each domain and also for the aggregate score. For PEDIS, ICC (1,1) and ICC (1,k) were reported only for each domain, with 95% confidence intervals.

Statistical analysis was performed using the "psych" package in R (version 3.1.3 (2015), R Foundation for Statistical Computing, Vienna, Austria).

When reporting ICC (1,1) and ICC (1,k), a result of 0 signified no agreement, whilst a result of 1 signified absolute agreement between observers. Whilst there is no absolute consensus on how to interpret the parameters of agreement between 0 and 1, the subjective guidelines provided for the kappa coefficient were reasonably applied, ie: 0.01 = poor; 0.01-0.2 = slight; 0.21-0.4 = fair; 0.41-0.6 = moderate; 0.61-0.8 = substantial; 0.81-1.00 = almost perfect.[10]

Sample size calculation

Our null hypothesis was that there was only a fair amount of agreement between raters i.e. an ICC of 0.3. We used the "ICC.Sample.Size" package to calculate the required sample size assuming only 3 raters per patient (with alpha = 0.05 and power = 0.8). Between 9 and 30 patients would be required

to identify substantial levels of agreement (ICC between 0.8 and 0.61). We therefore aimed to enrol at least 30 subjects into the study.

Results

Participants / observers

Some 45 patients were identified as potentially eligible (Figure 1). A total of 37 patients were included in the study and assessed by a pool of 12 observers. Some patients were assessed by more than observer in each specialty (e.g. two podiatrists) and therefore the totals in the ulcer column may exceed 37 (Table 2 and Table 3).

Outcomes

Single-observer observations

Reliability for single observers assessing individual patients (ICC (1,1)) was moderate when assessing overall UT and SINBAD scores (0.53 and 0.44, respectively). ICC (1,1) for PEDIS categories was fair to moderate and varied between 0.23 and 0.42. The worst agreement for single observers was when scoring infection (PEDIS 0.28; SINBAD 0.28), ischaemia (PEDIS 0.23, SINBAD 0.26) or both (UT 0.25) (Table 4).

Multiple-observer average ratings

Reliability for multiple observers' average ratings - ICC (1,k) - was almost perfect when assessing overall UT and SINBAD scores (0.94 and 0.91,

respectively), as well as individual categories in UT (0.81-0.94), SINBAD (0.82-0.99) and PEDIS scores (0.80-0.90). The best agreement for multiple observers was when scoring the site of ulceration (SINBAD 0.99) and the depth (UT 0.94, SINBAD 0.94). The worst agreement for multiple observers was when assessing ischaemia, however this was still considered to represent at least substantial agreement (SINBAD 0.82, PEDIS 0.80).

Discussion

This study demonstrates that the inter-observer reliability of SINBAD, PEDIS and UT wound classification systems is moderate at best when used by single observers assessing individual patients, and is particularly poor when assessing the important clinical parameters of infection and ischaemia. In contrast, there is almost perfect reliability when multiple observers (from a pool of observers) assess the same patients, particularly when assessing depth and site of the ulcer.

The worst agreement between multiple observers was achieved when assessing ischaemia, however this was still considered to represent substantial agreement. Infection and ischaemia have been shown to be important predictors of outcome in patients with DFU,[11, 12] however this study has demonstrated that the diagnosis of PAD and infection in clinical practice is challenging, even when using standardised scoring systems. Other scoring systems, such as the WIfI (Wound, Ischaemia and foot Infection) and IDSA (Infectious Diseases Society of America) systems deal more objectively with infection and ischaemia in DFU and are alternatives to those assessed in this study.

Whilst many studies have been adequately validated, there are few previous studies that report inter-observer reliability in DFU classification systems. It is important to recognise the difference between validity and reliability. Validity assesses whether a concept measures what it is intended to measure (in this case, factors that contribute to outcomes in DFU), whereas reliability deals with the overall consistency of the measurement. In one study investigating the use of the S(SA)SAD system, inter-observer reliability was reported to be 'good', however original data was not supplied.[13] The authors of the St Elian score reported a kappa coefficient of 0.61-1.00 when 2 observers independently classified the wounds.[14] A more recent study comparing the UT and Meggitt-Wagner systems, using digital photographs of DFU, found only moderate agreement amongst the group of clinicians, and significantly higher agreement between nurses than doctors.[15]

The wide variation in presentation, aetiology and outcomes of patients with DFU makes it difficult to select a single scoring system for widespread use, particularly as the prevalence of influencing factors, such as PAD, varies across the world and the factors most strongly associated with outcomes depend on the population studied.[16] In addition, whilst many of the well-known systems have been internally validated, there is a lack of robust external validation for many scores, as well as poor reliability when used on a global scale by different types of health professionals. [17, 18, 19]

This study has some important strengths. The use of a multi-disciplinary pool of observers should capture data from clinicians with a range of clinical expertise, training and should reflect standard practice in other centres. It also represents the fact that scoring systems should be designed for use by a

range of health care professionals. The use of *in vivo* wound assessment in the present study (rather than photographs) allowed assessment of many aspects of the scoring systems that would not be possible if photographs alone were assessed - it has previously been demonstrated that wound classification using photographs is limited. [20] The present study has demonstrated similarity in the reliability of the three scores and these results may therefore be cautiously extrapolated to other systems that assess the same domains. In addition, the use of the intra-class correlation statistic allowed analysis of both single- and multiple-observer reliability – which are both important aspects to evaluate when considering whether a scoring system may be useful for research, audit or clinical purposes.

However, the *in vivo* approach did not allow assessment of intra-observer variability, due to the potential for significant fluctuation in wound severity between sequential assessments and the requirement for assessor blinding, which would be difficult to achieve in the clinical setting. Also, some of the clinician assessors may have been familiar with the participants' prior medical history prior to enrolment - therefore, bias may have been introduced during the scoring assessments. However, this reflects a real-life situation where clinicians are often performing serial clinical assessments on patients well known to their team.

In addition, whilst the PEDIS score may include ABI, TBI and TcPO2 measurements for assessment of ischaemia, assessors in this study used only palpation of foot pulses. It could be noted, however, that the authors of the PEDIS system specify 'when resources are lacking the system could be easily adapted for local use'. [3] The use of palpation of pulses as the only

method of determining the presence or absence of ischaemia may be deemed inadequate in a country with adequate resources, however may represent current practice in some countries without the tools to perform objective testing. In this study, this represents a source of bias when considering the results of the ischaemia testing using the PEDIS score. It is therefore perhaps not surprising that the worst reliability when using the PEDIS score is when assessing ischaemia. However, the SINBAD and the UT scoring systems permit the use of pulse palpation when assessing ischaemia and the reliability for ischaemia testing was comparably low when using these scores.

There was also no comparison between the assessors' observations of neuropathy and the results of objective testing. Whilst this may be a weakness, the aim of this study was not to assess validity of the scoring systems but to assess inter-observer reliability and a gold standard was therefore not used.

This study has demonstrated that, when assessing patients with DFU using PEDIS, UT and SINBAD scoring systems, single-observer reliability is poor. This may reflect that such classification systems, even if they have been validated, may not be as useful in the clinical context (for example, when an individual clinician assesses an individual patient or uses the score to make a referral to another speciality). However, the reliability of multiple-observers' average ratings was almost perfect, which may justify the use of classification systems for research or audit purposes – for example, when multiple

observers provide average scores for a group of patients – and can justify their use to compare between centres.

As well as reporting internal and external validation data, all proposed scoring systems should report inter-observer reliability, in order to be accepted as a potentially useful tool for patients with DFU.

Funding: Nil

Conflicts of interest: The authors declare that they have no conflicts of interest

References

1. Reiber GE, Lipsky BA, Gibbons GW. The burden of diabetic foot ulcers. *Am J Surg*. 1998;176(2A Suppl):5S–10S.

2. Karthikesalingam A, Holt PJE, Moxey P, Jones KG, Thompson MM, Hinchliffe RJ. A systematic review of scoring systems for diabetic foot ulcers. *Diabet Med*. 2010;27(5):544-549..

3. Schaper NC. Diabetic foot ulcer classification system for research purposes: a progress report on criteria for including patients in research studies. *Diabetes Metab Res Rev*. 2004;20 Suppl 1(S1):S90-S95.

4. Ince P, Abbas ZG, Lutale JK, et al. Use of the SINBAD classification system and score in comparing outcome of foot ulcer management on three continents. *Dia Care*. 2008;31(5):964-967.

5. Lavery LA, Armstrong DG, Harkless LB. Classification of diabetic foot wounds. *J Foot Ankle Surg*. 1996;35(6):528-531.

6. National Institute for Health, Excellence C. *NICE Guidelines [CG119] Diabetic Foot Problems*. 2011.

7. Shrout PE, Fleiss JL. Intraclass Correlations - Uses in Assessing Rater Reliability. *Psychol Bull*. 1979;86(2):420-428.

8. Rousson V. Assessing inter-rater reliability when the raters are fixed: Two concepts and two estimates. *Biom J*. 2011;53(3):477-490.

9. Carrasco JL, Jover L. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*. 2003;59(4):849-858.

10. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33(1):159.

11. Jude EB, Oyibo SO, Chalmers N, Boulton AJ. Peripheral arterial disease in diabetic and nondiabetic patients: a comparison of severity and outcome. *Dia Care*. 2001;24(8):1433-1437.

12. Prompers L, Schaper N, Apelqvist J, et al. Prediction of outcome in individuals with diabetic foot ulcers: focus on the differences between individuals with and without peripheral arterial disease. The EURODIALE Study. *Diabetologia*. 2008;51(5):747-755.

13. Treece KA, Macfarlane RM, Pound N, Game FL, Jeffcoate WJ. Validation of a system of foot ulcer classification in diabetes mellitus. *Diabet Med*. 2004;21(9):987-991.

14. Jesús FRM-D. A Checklist System to Score Healing Progress of Diabetic Foot Ulcers. *Int J Low Extrem Wounds*. 2010;9(2):74-83.

15. Santema TB, Lenselink EA, Balm R, Ubbink DT. Comparing the Meggitt-Wagner and the University of Texas wound classification systems for diabetic foot ulcers: inter-observer analyses. *Int Wound J*. February 2015:[Epub ahead of print].

16. Abbas ZG, Lutale JK, Game FL, Jeffcoate WJ. Comparison of four systems of classification of diabetic foot ulcers in Tanzania. *Diabet Med.* 2008;25(2):134-137.

17. Fischer M, Rüegg S, Czaplinski A, Strohmeier M. Research Inter-rater reliability of the Full Outline of UnResponsiveness score and the Glasgow Coma Scale in critically ill patients: a prospective observational. *Crit Care*. 2010;14(2):R64.

18. Gill M, Martens K, Lynch EL, Salih A, Green SM. Interrater reliability of 3 simplified neurologic scales applied to adults presenting to the emergency department with altered levels of consciousness. *Ann Emerg Med*. 2007;49(4):403-407.

19. Monteiro-Soares M, Martins-Mendes D, Vaz-Carneiro A, Sampaio S, Dinis-Ribeiro M. Classification systems for lower extremity amputation prediction in subjects with active diabetic foot ulcer: a systematic review and meta-analysis. *Diabetes Metab Res Rev*. 2014;30(7):610-22

20. Murphy RXJ, Bain MA, Wasser TE, Wilson E, Okunski WJ. The Reliability of Digital Imaging in the Remote Assessment of Wounds:

Defining a Standard. *Ann Plast Surg*. 2006;56(4):431-436.